STREAM3D: SEQUENTIAL MULTI-VIEW 3D GENERATION VIA EVIDENTIAL MEMORY

Anonymous authors
Paper under double-blind review

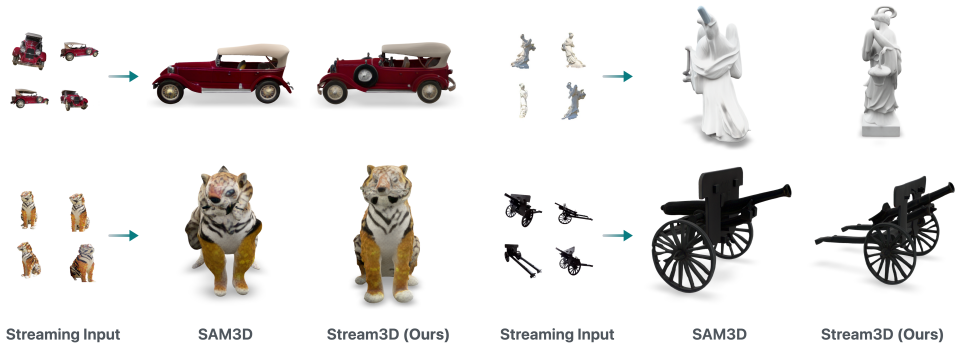


Figure 1: **Stream3D** takes streaming input views as additional conditioning signals to improve the performance of pretrained single-view-conditioned 3D generation models without retraining pretrained weights. Compared with SAM-3D, incorporating views from the input stream can substantially improve 3D generation quality.

ABSTRACT

View-conditioned 3D generators such as SAM 3D, TRELLIS and Hunyuan3D produce high-quality object 3D representations from a single view, but real-world visual observation often arrives as long monocular streams. Naively applying these generators to each streaming frame independently leads to severe temporal inconsistency in the generated results. To address this problem, we propose Stream3D, the first training-free streaming mechanism that turns a frozen view-conditioned 3D generator into a streaming generator with constant cross-chunk evidential memory. Stream3D achieves this by maintaining a compact evidential memory, which selectively caches the most informative historical frames based on a proposed evidence score mechanism. As the stream progresses, the memory dynamically updates to retain a fixed number of informative frames, preventing the evidence cache from growing linearly with sequence length. This sustains generation quality over long sequences and keeps the underlying generator completely unchanged without retraining, architectural modifications, or auxiliary losses. Evaluated on both realistic and synthetic streaming benchmarks, Stream3D outperforms latent-transport baselines, including KV-cache reuse and flow-based feature editing, across both photometric and geometric metrics. More details can be found at: Link.

1 INTRODUCTION

Object-centric 3D generation is becoming a practical building block for vision and robotics (Liu et al., 2023a; Li et al., 2023; Tang et al., 2024a). Recent systems such as SAM 3D Objects (Chen et al., 2025b) and TRELLIS.2 (Xiang et al., 2025a) can take an image and reconstruct corresponding Gaussian splat (Kerbl et al., 2023) or 3D mesh within seconds. However real capture devices produce long monocular streams: a phone circling an object (Mildenhall et al., 2020) or a robot observing a scene while moving (Huang et al., 2024). Naively applying a single-view 3D generator frame by frame yields temporally inconsistent reconstructions and exploits only partial observations (Chen et al., 2025b; Xiang et al., 2025a). Feeding all frames jointly through multi-diffusion (Bar-Tal et al., 2023) or multiview fusion (Li et al., 2026) is computationally expensive and can become infeasible

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for long streams, while processing fixed-size chunks with flow-matching-based editing (Kulikov et al., 2025) discards historical context needed for global consistency.

An alternative approach is to adapt techniques from streaming video generation or 3D reconstruction (Kim et al., 2024; Henschel et al., 2024; Lan et al., 2025; Zhuo et al., 2026). For instance, state-transport mechanisms like KV banks (Lan et al., 2025) or FlowEdit-style velocity edits (Kulikov et al., 2025) could be introduced to propagate information across chunks. However, transporting latent states in this manner is also problematic. As the orientation, shape, and scale of these states are deeply entangled within the generator, they are difficult to align across frames, leading to severe error accumulation during the streaming process (Li et al., 2026). Furthermore, as these methods transport and accumulate state over time, their memory footprint naturally grows with sequence length (Chen et al., 2026). Such state propagation may also introduce compounding errors over long streams, echoing the error accumulation and memory bottlenecks observed in autoregressive video diffusion models (Wang et al., 2025b). Thus, the mechanism intended to preserve consistency can become a source of degradation when the transported state is misaligned or stale.

To solve this problem, we propose **Stream3D**, a training-free streaming mechanism that turns a frozen view-conditioned 3D generator into a long-stream generator without modifying its weights or architecture. Rather than transporting latent states across chunks, Stream3D retains *evidential views*: observed frames that the generator itself identifies as reliable conditioning signals through its cross-attention maps. Specifically, during a lightweight warmup pass, we compute an evidence score for each query token and incoming views, measuring whether the token attends to that views both strongly and selectively. These scores update a token-level *Adaptive Evidential Memory*, which stores only a fixed number of high-evidence frame indices per query token, keeping this memory constant with stream length. This token-level design allows different spatial regions of the 3D volume to retrieve different historical observations, rather than relying on a single global frame subset or an accumulated latent state. As new views arrive, the retained evidence for each query token can only be maintained or improved.For generation, token-level evidence is aggregated into frame-level ownership scores, and the top-K frames are selected as a bounded conditioning bundle for *Evidence-Based Multi-Generation*. In this way, Stream3D avoids latent-space alignment, preserves long-range visual evidence, and enables coherent streaming 3D generation while leaving the original generator unchanged. We evaluate Stream3D as a lightweight mechanism around pre-trained SAM 3D on long monocular streams as shown in Fig. 1. Our contributions are fourfold:

- **Streaming 3D generation.** We pioneer the task of extending frozen view-conditioned 3D generators to long monocular streams, supporting temporally consistent 3D generations while keeping evidential memory bounded and avoiding retraining.
- **Adaptive Evidential Memory.** We introduce a compact memory mechanism that stores token-level evidential views rather than transporting latent states. The evidence cache is training-free and remains constant in size with stream length.
- **Evidence-Based Multi-Generation.** We propose an evidence-guided generation strategy that aggregates token-level memory into a bounded conditioning bundle and runs the frozen generator on the selected views. This enables different spatial regions to draw from the more reliable historical observations without modifying the underlying generator.
- **Superior streaming performance.** We show that Stream3D outperforms latent-transport and fixed-view baselines across photometric and geometric metrics, while sustaining generation quality over long streams with a constant evidential-memory footprint.

2 Related Work

2.1 3D Generation

Recent object-centric 3D generators can be understood along three main design axes: *output representation*, *learning regime*, and *input format*. Along the first axis, different methods adopt different 3D representations, including triplane- and Neural Radiance Field (NeRF)-based backbones (Hong et al., 2023; Tochilkin et al., 2024; Mildenhall et al., 2020; Yu et al., 2021; Chan et al., 2022; Li et al., 2023; 3DTopia, 2023; Wang et al., 2023b), surface meshes (Xu et al., 2024a; Wu et al., 2024a; Xiang et al., 2025b; Liu et al., 2023a; Wang et al., 2024b; Wei et al., 2024; Tang et al., 2023; Gao

et al., 2022), 3D Gaussian splats (Huang et al., 2025; Kerbl et al., 2023; Tang et al., 2024a; Xu et al., 2024b; Tang et al., 2024b; Yi et al., 2024), and native structured latent volumes (Zhao et al., 2025; Wu et al., 2024b; Li et al., 2025; 2024b; Chen et al., 2025b; Xiang et al., 2025a). Along the second axis, existing approaches cover closed-form regressors (Hong et al., 2023; Tochilkin et al., 2024; Xu et al., 2024a; Wu et al., 2024a; Huang et al., 2025; Xiang et al., 2025a; Li et al., 2023; Liu et al., 2023a; Tang et al., 2024a; Wang et al., 2024b; Wei et al., 2024; 3DTopia, 2023; Xu et al., 2024b; Wang et al., 2023b), 3D latent diffusion models (Zhao et al., 2025; Wu et al., 2024b; Li et al., 2025; 2024b; Chen et al., 2025b; Nichol et al., 2022; Jun & Nichol, 2023), and SDS-based optimization methods (Poole et al., 2022; Lin et al., 2023; Wang et al., 2023c;a; Metzer et al., 2023; Chen et al., 2023; Tang et al., 2023; Qian et al., 2023; Tang et al., 2024b; Yi et al., 2024; Sun et al., 2023; Qiu et al., 2024). Some layout-aware variants (Chen et al., 2025b; Li et al., 2026; Anciukevičius et al., 2023) further model the spatial arrangement of multiple objects within a scene. The third axis, *input format*, is the one most relevant to our work. Despite their differences in representation and learning paradigm, most existing methods assume a fixed and limited input interface: they take either a single image or a small, predefined set of images as input. Multi-view diffusion methods (Liu et al., 2023b; Shi et al., 2023a; Liu et al., 2023c; Long et al., 2024; Shi et al., 2023b; Kong et al., 2024; Voleti et al., 2024; Li et al., 2024a) relax this setting by first hallucinating additional views before reconstructing 3D geometry. Video-conditioned reconstruction models (Wang et al., 2024a; Leroy et al., 2024; Yang et al., 2025; Zhou et al., 2026b; Wang et al., 2025a; Chen et al., 2024b) and 4D-aware generators (Bahmani et al., 2024; Singer et al., 2023; Zheng et al., 2023; Zhao et al., 2023; Jiang et al., 2024; Chen et al., 2024a; Ren et al., 2024; Zhou et al., 2026a) extend the input format further to short temporal clips. More recently, MV-SAM3D (Li et al., 2026) enables multi-view fusion directly in the latent space of a 3D generator. Nevertheless, these methods still operate on a fixed input bundle determined in advance, rather than supporting truly open-ended streaming inputs. None of these methods natively supports unbounded online streams. In practice, long-stream processing is typically handled through ad hoc latent-transport schemes, which propagate intermediate states across chunks. However, such designs usually incur memory growth with sequence length and lead to performance degradation over long horizons. In this work, we adapt a frozen view-conditioned generator (Chen et al., 2025b) into a streaming generator that maintains constant cross-chunk evidential memory, independent of stream length, and agnostic to the model backbone.

2.2 Reconstruction from Streaming Inputs

Reconstructing 3D geometry from streaming inputs has traditionally been studied in monocular SLAM (Engel et al., 2014; 2018; Davison et al., 2007; Klein & Murray, 2007; Newcombe et al., 2011; Forster et al., 2014; Mur-Artal et al., 2015; Mur-Artal & Tardós, 2017; Campos et al., 2021; Sun et al., 2021; Matsuki et al., 2024; Huang et al., 2024; Zhou et al., 2025), which incrementally estimates camera motion and scene structure from video. Recent methods have extended modern feed-forward reconstruction models to online settings. Notably, Spann3R (Wang & Agapito, 2025) augments a DUSt3R-style encoder (Wang et al., 2024a) with a token-addressable spatial memory, enabling online pointmap fusion over long image streams. Similarly, SLAM3R (Liu et al., 2025), also built upon DUSt3R, introduces a real-time end-to-end dense reconstruction system that directly predicts 3D pointmaps from RGB videos. Point3R (Wu et al., 2025) further incorporates an explicit geometry-aligned spatial pointer memory, together with 3D hierarchical RoPE and an adaptive fusion mechanism. However, DUSt3R itself remains inherently two-view, restricting each inference step to a fixed image pair and making large-scale fusion dependent on iterative matching and optimization. VGGT-SLAM (Maggio et al., 2025; Wang et al., 2025a) addresses this limitation by adopting the more powerful VGGT transformer, which supports image sets of arbitrary length. CUT3R (Wang et al., 2025c) instead adopts an RNN-style formulation for causal pointmap prediction from unstructured image streams. However, it compresses all past observations into a limited recurrent state, which can hinder long-range memorization and fine-grained multi-view fusion. Following the design philosophy of modern large language models, StreamVGGT (Zhuo et al., 2026) and Stream3R (Lan et al., 2025) employ causal transformers to implicitly cache historical visual tokens. In addition, because CUT3R suffers from severe drift on long streaming inputs, TTT3R (Chen et al., 2025a) proposes a simple empirical state-update rule to improve sequence-length generalization. Stream3D departs from these reconstruction-centered approaches: while online reconstruction focuses on aggregating geometry already observed in the input stream, streaming 3D generation must also infer and synthesize unseen structure under temporal and geometric consistency constraints.

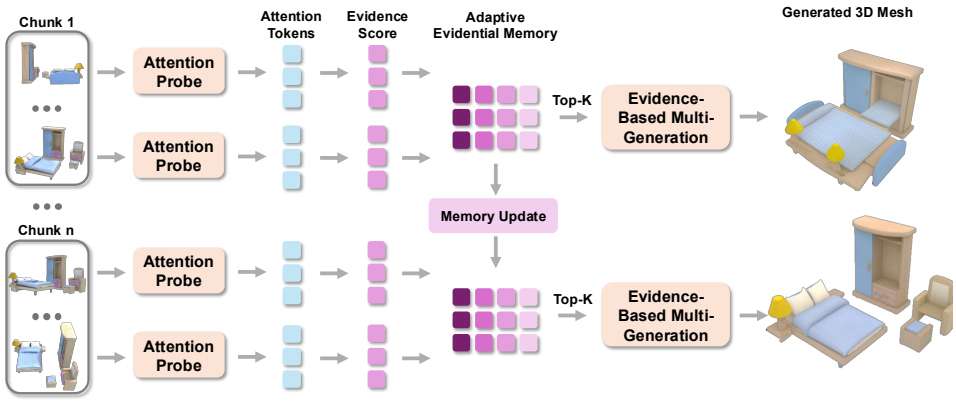


Figure 2: **Framework of Stream3D.** Given a streaming video, Stream3D processes frames chunk by chunk. A lightweight warmup pass extracts token-wise evidence score from cross-attention, which is stored to vote for informative frames to update the evidential memory, i.e., **Adaptive Evidential memory**. The selected frames condition the structure stage, after which views are selected for the appearance stage on the resulting sparse structure. Stream3D maintains a fixed-capacity token-level evidence cache and selects a bounded conditioning bundle.

3 METHOD

We extend view-conditioned 3D generators, SAM 3D Objects (Chen et al., 2025b) and TRELLIS.2 (Xiang et al., 2025a) — to handle long streaming inputs without retraining shown in Fig. 2. The *Adaptive Evidential Memory* (Sec. 3.2) enables efficient long-range memory by retaining token-level evidence from past views, while *Evidence-Based Multi-Generation* (Sec. 3.3) leverages this memory to select and fuse historical conditioning views for each reconstruction.

3.1 PROBLEM SETUP

3D Generation Preliminary. Let f_θ denote a frozen view-conditioned 3D generator that, given a single input frame v and an initial noise prior $z_0 \sim \mathcal{N}(0, I)$, produces a 3D sample (Gaussian splat (Kerbl et al., 2023), mesh, or latent volume) $\hat{y} = f_\theta(v; z_0)$. Recent 3D generation models, i.e., SAM 3D (Chen et al., 2025b), TRELLIS (Xiang et al., 2025b), TRELLIS.2 (Xiang et al., 2025a), Hunyuan3D 2.0 (Zhao et al., 2025) and CraftsMan3D (Li et al., 2024b), instantiate f_θ as a two-stage pipeline — a structure stage (SS) producing a coarse occupancy / latent grid, followed by a texture or appearance stage (SLAT) — with at least one cross-attention layer of the form:

$$\mathbf{A}_v = \text{softmax}\left(\frac{\mathbf{Q}_v \mathbf{K}_v^\top}{\sqrt{d}}\right)_{:,\mathcal{I}_v} \in [0,1]^{Q\times P}, \tag{1}$$

Cross-Attention Map; Conditioning Keys of View v; Attention Shape; Query Tokens for View v; Scale Factor

where Q is the number of query tokens in the generator's voxel grid (e.g., $16^3 = 4096$ in SAM 3D's structure stage), $\mathbf{Q}_v$ is the query representation in view v's forward pass, $\mathbf{K}_v$ contains that view's image and other conditioning keys (including point-map keys when used), $\mathcal{I}_v$ indexes its P image-patch keys, and d is the feature dimension of each query/key token. Softmax is applied over the keys of view v before extracting the patch attention $\mathbf{A}_v$; keys from other views are not included. Batch and attention-head indices are omitted for clarity. In its native form, 3D generator f_θ accepts a single frame and produces a single reconstruction; it has no mechanism for incorporating multi-view evidence.

Streaming 3D Generation. Our method extends 3D generator f_θ to accept a stream of condition views $\mathcal{V} = \{v_1, \ldots, v_T\}$ by fusing the per-view forward passes through a confidence-weighted Multi-Diffusion-style aggregation in 3D latent space. The stream is partitioned into consecutive chunks $\mathcal{C}_k$ of size C. At chunk k, we update the 3D sample $\hat{y}_k$ using evidence from newly observed and historical views, with a fixed-capacity evidence cache and a bounded conditioning bundle.

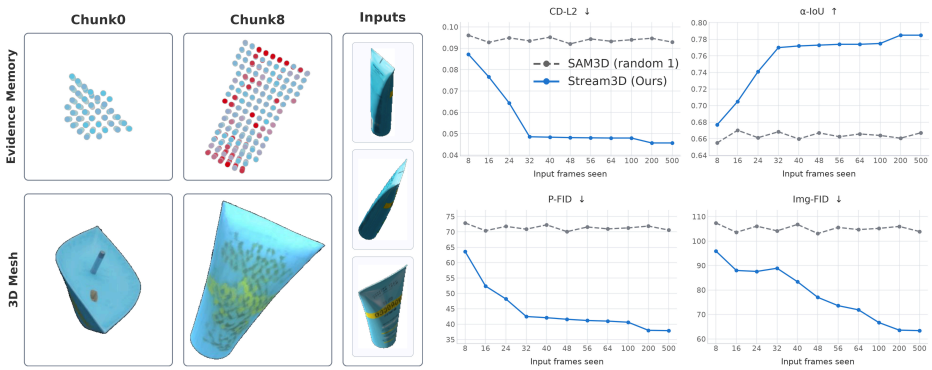


Figure 3: **Adaptive Evidential Memory.** Left: Given a streaming sequence, our method updates a compact token-level memory by retaining historical views with the strongest evidence for each query token. The color transition from blue to red denotes increasing frame indices, from earlier to later observations. Over time, the memory incorporates evidence from increasingly diverse viewpoints, and the reconstruction improves from early to later chunks, consistent with the long-horizon geometry and appearance curves. Right: Compared with SAM3D, Stream3D improves as input frames accumulate through successive chunks and then reaches a stable plateau.

3.2 Adaptive Evidential Memory

To keep the cross-chunk evidence cache independent of the total sequence length T while preserving generation quality over long streams, we introduce an *Adaptive Evidential Memory* mechanism, shown on the left of Fig. 3. The right side of Fig. 3 shows that Stream3D improves as more frames are observed and then reaches a stable performance plateau. Appendix B.1 analyzes geometry and appearance across chunks. Rather than blindly retaining all historical frames, this approach dynamically filters and preserves only the most informative views by evaluating their relevance at a per-token level. Our mechanism operates in three distinct stages. (1) First, we compute a token-wise **Evidence Score** via a lightweight attention probe to measure the significance of each incoming view. (2) Second, we update a fixed-capacity global **Memory** that persistently tracks the highest-scoring frames for each individual query token. (3) Finally, we perform **Conditioning-view selection** by allowing the tokens to "vote" for their preferred frames, aggregating these local preferences to select an bounded-size condition set for the full generation pass.

Evidence Score. At each chunk, we probe each newly observed view with one flow/denoising step of the structure-stage generator, using a *frozen* prior z_0 sampled once at chunk 0 and reused for all subsequent chunks. Views can be processed in a batch, but their conditioning and attention computations are independent. From a fixed cross-attention layer L, we extract the cross-attention map in Eq. equation 1. For each non-special query token q and each view v, we compute a per-token evidence score $\mathbf{M}_v[q]$ that measures how informative view v is for reconstructing token q.

The evidence score should capture two complementary properties: the share of view v's conditioning attention assigned to its image patches and how spatially concentrated that attention is within those patches. We therefore define a normalized attention distribution over the P patch tokens of view v:

$$\mathbf{H}_v[q] = -\frac{1}{\log(P)} \sum_{i=1}^{P} \tilde{\mathbf{A}}_v[q,i] \log \tilde{\mathbf{A}}_v[q,i], \qquad \tilde{\mathbf{A}}_v[q,i] = \frac{\mathbf{A}_v[q,i]}{\sum_{j=1}^{P} \mathbf{A}_v[q,j]}. \tag{2}$$

Here, $\mathbf{H}_v[q] \in [0,1]$ is the normalized entropy of the attention distribution, where lower entropy indicates that the query token attends to a more localized set of patches. Let $s_v[q] = \sum_{i=1}^{P} \mathbf{A}_v[q,i]$ denote the patch attention mass. We then define the evidence score as

$$\mathbf{M}_v[q] = \frac{1+\beta s_v[q]}{1+\beta} \cdot (1 - \mathbf{H}_v[q]), \tag{3}$$

where $\beta = 8$ controls the contribution of patch attention mass. Equations equation 1–equation 3 suppress batch and head indices: patch mass and entropy confidence are computed per head and averaged separately over batch and heads before their product is formed. Since $s_v[q] \in [0,1]$ and $\mathbf{H}_v[q] \in [0,1]$, the evidence score lies in $[0,1]$. The fixed divisor $1+\beta$ places scores from different views and chunks on the same scale without using a statistic of the current probe.

This design follows a mass–concentration decomposition. The mass $s_v[q]$ is the share of attention allocated to image patches among view v's own conditioning keys; $1 - \mathbf{H}_v[q]$ rewards spatially concentrated attention within those patches. High-entropy attention spread over many patches is treated as weak evidence even when image-patch mass is large, since it may correspond to diffuse background response or non-specific visual context. Because the one-step probe evaluates each view separately from the same z_0, a fixed view's score does not depend on the other views assigned to its chunk, provided its input conditions and probe settings are unchanged. This gives recorded scores a common numerical basis for cross-chunk top-D comparison; it does not make the score a calibrated probability of reconstruction correctness.

Memory. The cross-chunk evidence memory is represented by two $Q \times D$ arrays, $\mathbf{M}$ and $\mathbf{F}$, where Q is the number of query tokens and D is the memory depth. Historical observations are read on demand from external storage. $\mathbf{M} \in \mathbb{R}^{Q \times D}$ stores the D token-wise highest evidence scores over all observed views, while $\mathbf{F} \in \mathbb{Z}^{Q \times D}$ stores their corresponding global frame indices. During the streaming process, each token's top-D list is updated by merging in the new candidates from newly arriving frames. Only the top-D entries for each token are retained in this cache.

Conditioning-view Selection. At a certain chunk, the runtime view set $\mathcal{V}^\star$ selected for the full forward pass is obtained by aggregating the per-token top-D lists into a per-frame *token-ownership count*, then taking the **top-K** frames by that count. Concretely, we define the ownership count of frame f_r as the number of (token, rank) slots it occupies anywhere in $\mathbf{F}$:

$$n_{f_r} \;=\; \big|\{(q, j) : \mathbf{F}[q, j] = f_r,\; 1 \le q \le Q,\; 1 \le j \le D\}\big|, \tag{4}$$

and select the K frames with the largest counts for $\mathcal{V}^\star$. Notably, there are two distinct ranks D and K. D is the *memory depth*: how many candidate frames each token retains in its sorted list; K is the *bundle size*: how many frames the downstream generator consumes per forward pass. Here, D controls how robust the evidence score computation is.

3.3 Evidence-Based Multi-Generation

Generation via multi-view flow-matching fusion. The evidential memory selects the structure-stage (SS) conditioning bundle. After SS produces the current sparse structure, we probe the observed views independently on that structure at the appearance stage (SLAT), score them with Eq. equation 3, and count which view has the highest score for each SLAT latent. These counts determine a bounded SLAT conditioning bundle. For either stage, let $\mathcal{V}^\star$ denote its selected views. At each flow step t, the frozen generator evaluates a velocity $V_\theta(z_t, v)$ for every $v \in \mathcal{V}^\star$ from the same latent z_t and fuses the velocities into one update. The fused velocity at query token q is a view-weighted average:

$$\bar{V}_\theta(z_t)[q] \;=\; \sum_{v \in \mathcal{V}^\star} \overline{M}_v[q]\, V_\theta(z_t, v)[q], \qquad \sum_{v \in \mathcal{V}^\star} \overline{M}_v[q] = 1, \tag{5}$$

where $\overline{M}_v[q]$ is obtained by recomputing the per-view score in Eq. equation 3 on the selected conditioning views and normalizing it across $\mathcal{V}^\star$ for each query token. The scores used for SS fusion and SLAT fusion are computed separately at their respective attention layers. The full forward pass is $\hat{y} = f_\theta(\mathcal{V}^\star; z_k)$. At every chunk, we reset the seed to the same fixed value before initializing z_k.

Algorithm and properties. Algorithm 1 in Appendix A summarizes the per-chunk procedure. The method uses attention layers for SS and SLAT, memory depth D, bundle size K, score coefficient β, a fixed prior z_0, and no learned parameters. Processing each chunk requires a one-step SS probe of newly observed views, a memory update, and SS and SLAT generation with their selected conditioning bundles.

Summary. The Adaptive Evidential Memory has two structural properties that clearly distinguish it from latent-transport schemes. First, its cross-chunk memory footprint is bounded by at most $2 \times Q \times D$ scalars, and is therefore constant with respect to the stream length T. Second, for every query token q and rank j, the cached evidence score $\mathbf{M}[q, j]$ is monotonically non-decreasing over time, since each update retains only the top-D entries from the union of the previous cache and the newly arrived candidates. The SS conditioning views are then selected by aggregating token-wise ownership counts, while the evidence-memory cost remains fixed. For a fixed set of observed frames, per-view inputs, and probe prior, the SS scores in Eq. 3 and the resulting top-D memory are invariant to how those frames are partitioned into chunks. Thus the SS ownership counts and their top-K selection are also unchanged at a matched observation horizon.



Figure 4: **Qualitative results on GSO and NAVI.** Stream3D produces more consistent and geometrically faithful 3D generations than single-view and multi-view diffusion baselines.

Table 1: **Main results on GSO and NAVI.** Stream3D consistently improves geometry and appearance over single-view and multi-view generation baselines.

Data	Method	Geometry			Appearance			
		CD↓	IOU↑	PFID↓	PSNR↑	SSIM↑	LPIPS↓	Image FID↓
GSO	StreamVGGT	0.064	0.549	75.520	10.800	0.787	0.258	189.180
	EscherNet + NeuS	0.063	0.664	49.570	14.790	0.840	0.151	120.990
	TRELLIS.2	0.146	0.433	140.513	12.281	0.830	0.201	141.006
	TRELLIS+M.D.	0.053	0.750	44.515	15.933	0.854	0.149	**60.091**
	TRELLIS.2+M.D.	0.093	0.648	87.720	14.082	0.839	0.176	98.255
	SAM3D	0.094	0.664	71.263	14.178	0.848	0.178	105.197
	Stream3D(TRELLIS.2)	0.087	0.692	77.730	14.213	0.845	0.169	83.279
	Stream3D(SAM3D)	**0.048**	**0.775**	**40.641**	**16.145**	**0.866**	**0.139**	66.711
NAVI	TRELLIS.2	0.071	0.818	42.045	20.362	0.934	0.075	88.067
	TRELLIS+M.D.	0.055	0.810	44.297	20.426	0.935	0.078	87.654
	TRELLIS.2+M.D.	0.066	0.791	50.839	19.914	0.934	0.078	89.755
	SAM3D	0.068	0.799	47.989	19.445	0.930	0.086	89.593
	Stream3D(TRELLIS.2)	0.045	**0.833**	**40.431**	**20.880**	0.936	**0.072**	85.546
	Stream3D(SAM3D)	**0.043**	0.825	42.151	20.720	**0.939**	0.073	**81.173**

4 EXPERIMENT

We evaluate Stream3D on long-stream 3D generation, where the model receives continuously arriving posed images and generates a 3D reconstruction for each chunk. We compare reconstruction quality with single-view, multi-view, and streaming baselines, and analyze the effects of memory size and view-selection strategy.

4.1 EXPERIMENTAL SETUP

Implementation: Our experiments are run on NVIDIA H100 GPU. We use SAM3D with its standard settings as the underlying generation backbone. We use Depth Anything 3 (Lin et al., 2026) to predict monocular depth independently for each input frame and to estimate camera poses for point-map conditioning. Images arrive in video order and are processed by Stream3D in chunks of 8 frames. We set $K = 8$ and $D = 5$ for efficiency, and provide ablation studies to justify this choice. Fig. 3 uses additional 500-frame streams rendered from GSO GT 3D assets to examine long-horizon behavior.

Dataset: We evaluate on two complementary benchmarks. The first is the GSO benchmark (Downs et al., 2022), which contains high-quality scanned objects with ground-truth 3D assets. Following prior multi-view generation and reconstruction protocols, we render posed image streams from each object and evaluate both novel-view appearance and geometric accuracy. This controlled setting allows us to measure whether the generated 3D content remains faithful to the underlying object as the stream length increases. The second benchmark is NAVI (Jampani et al., 2023), which provides object-centric videos with varied camera trajectories.

Baseline: We compare Stream3D with representative single-view, multi-view, and streaming baselines. Single-view baselines, including SAM3D, and TRELLIS, generate 3D content from individual observations and provide reference results for evaluating the benefit of additional views.EscherNet (Kong

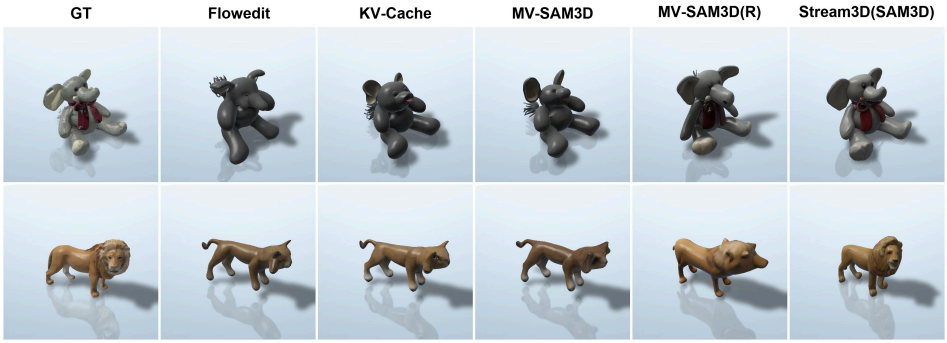


Figure 5: **Qualitative result of our Ablation studies.** FlowEdit denotes SAM3D with FlowEdit, and KV-Cache denotes SAM3D with KV-cache reuse. MV-SAM3D denotes MV-SAM3D applied to the last input chunk, while MV-SAM3D(R) denotes MV-SAM3D with K randomly selected views.

Table 2: **Ablation studies of different streaming strategies.**

Data	Method	Geometry			Appearance			
		CD↓	IOU↑	PFID↓	PSNR↑	SSIM↑	LPIPS↓	Image FID↓
GSO	SAM3D + FlowEdit	0.090	0.668	76.445	14.343	0.850	0.178	98.643
	SAM3D + KV-Cache	0.084	0.682	67.292	14.482	0.852	0.171	83.353
	MV-SAM3D + Last Chunk	0.113	0.630	86.044	13.906	0.848	0.184	104.228
	MV-SAM3D with K random views	0.064	0.676	68.534	14.828	0.859	0.156	83.039
	MV-SAM3D + Visibility	0.063	0.716	58.881	*15.05*	*0.855*	*0.163*	81.219
	MV-SAM3D + VLM	0.059	0.710	47.990	*15.23*	*0.855*	*0.157*	78.530
	MV-SAM3D + Stream3D	0.053	0.760	47.260	15.892	0.864	0.1449	77.750
	Ours (Fast)	0.051	0.759	43.940	15.510	0.861	0.152	80.610
	Ours	**0.048**	**0.775**	**40.641**	**16.145**	**0.866**	**0.139**	**66.711**

et al., 2024) serves as a strong multi-view baseline that benefits from multiple posed observations but is not designed for unbounded streaming inputs. StreamVGGT (Zhuo et al., 2026) serves as a reconstruction baseline with streaming input. We also compare against TRELLIS+M.D. (Xiang et al., 2025b) and TRELLIS.2+M.D. (Xiang et al., 2025a) which aggregate multiple views through full-context or multi-window diffusion but becomes expensive as the number of frames grows.

Metrics. We report appearance and geometry metrics. For appearance, we use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID) on held-out novel views. These metrics evaluate whether the generated representation renders images that are both photometrically accurate and perceptually faithful. For geometry, we report Chamfer Distance (CD), alpha-mask Intersection over Union (IoU), and point-cloud feature Fréchet distance (P-FID). These metrics measure geometric distance, foreground overlap, and the distribution of point-cloud features. Metric definitions and the evaluation protocol are provided in Appendix Sec. A.

4.2 Main Results

Table 1 reports results on GSO and NAVI, and Fig. 4 shows qualitative results on GSO. On GSO, Stream3D(SAM3D) obtains the best CD, IoU, P-FID, PSNR, SSIM, and LPIPS, while remaining competitive on Image FID. On NAVI, Stream3D achieves the best CD and Image FID, with consistently strong appearance scores. These results indicate improvements in both rendering quality and 3D structure. Compared with single-view baselines such as SAM3D and TRELLIS.2, Stream3D benefits from additional historical observations. It also improves over multi-view baselines such as TRELLIS+M.D. and TRELLIS.2+M.D. by retaining token-level evidence and selecting a bounded conditioning bundle. Across GSO checkpoints, CD and IoU stabilize after chunk 4 while P-FID and Image FID continue to improve, and the visualized reconstructions retain their overall geometry as surface details sharpen (Appendix B.1).

4.3 Streaming Baseline Comparison

Table 2 compares Stream3D with several streaming alternatives, and Fig. 5 reports qualitative results. (1) Stream3D improves over KV-cache reuse and FlowEdit. KV-cache reuse efficiently carries information across chunks, but cached states are not explicitly filtered by geometric reliability and can accumulate stale or ambiguous evidence. FlowEdit operates on fixed-size chunks and loses long-range

Table 3: **Ablation studies on module selection and hyperparameters.** "Layer" denotes the evidence extraction layers. "E" denotes the evidence weighting strategy. "D" denotes the memory depth. "K" denotes the number of selected Top-K views. Our setting is highlighted.

Layer	E	D	K	Geometry			Appearance			
				CD-L2↓	α-IoU↑	P-FID↓	PSNR↑	SSIM↑	LPIPS↓	Img-FID↓
SS6-SLAT3	Evi	5	8	0.057	0.744	48.62	15.577	0.859	0.1531	72.16
SS6-SLAT6	Evi	5	8	0.057	0.742	49.09	15.536	0.858	0.1528	70.97
SS9-SLAT3	Evi	5	8	0.048	0.775	41.36	16.136	0.866	0.1406	68.28
SS9-SLAT6	Evi	5	8	0.048	0.775	40.64	16.145	0.866	0.1396	66.71
SS9-SLAT9	Evi	5	8	0.048	0.776	41.05	16.119	0.865	0.1403	67.88
SS9-SLAT6	Evi(U)	5	8	0.056	0.749	44.98	15.730	0.863	0.1486	74.08
SS9-SLAT6	ENT	5	8	0.053	0.760	47.26	15.892	0.864	0.1449	77.75
SS9-SLAT6	Evi	1	8	0.051	0.769	41.15	15.952	0.866	0.1406	67.93
SS9-SLAT6	Evi	3	8	0.049	0.773	**39.47**	16.067	0.866	0.1415	66.45
SS9-SLAT6	Evi	7	8	0.049	0.772	40.19	16.040	0.863	0.1425	67.50
SS9-SLAT6	Evi	5	4	0.055	0.755	45.47	15.741	0.863	0.1478	70.20
SS9-SLAT6	Evi	5	12	0.048	0.775	40.12	16.167	0.865	0.1411	66.29
SS9-SLAT6	Evi	5	16	**0.048**	**0.779**	41.48	**16.169**	**0.867**	**0.1393**	**65.69**

observations outside the editing window. In contrast, Stream3D keeps a compact persistent memory and updates it according to token-level evidence. (2) Visibility- and VLM-based selection improve over random sampling, while token-level evidential selection provides further gains by retaining observations for individual spatial regions. (3) MV-SAM3D+Stream3D uses our selected views with MV-SAM3D fusion. Its improvement over other MV-SAM3D variants shows that evidential selection is useful by itself, while the remaining gap to Stream3D shows that evidence-weighted fusion also matters. (4) Ours (Fast), which applies Stream3D to the shortcut version of SAM3D, also achieves strong performance.

4.4 Ablation Studies

We conduct ablation studies in Tab. 3 to analyze the effects of attention layer,evidence weighting, memory depth, and TopK view selection. First, we study which transformer layers should be used for evidence computation. Following (Chen et al., 2025b), we denote the first generation stage as SS and the second as SLAT. Deeper layers generally yield better reconstruction performance, with SS performing well at layer 9 and SLAT at layer 6. Among the tested combinations, SS9-SLAT6 achieves strong geometry and appearance performance, suggesting that later transformer features provide better multi-view correspondence for streaming view selection. Second, we compare evidence weighting variants. Our product-form evidence weighting outperforms entropy-based weighting (ENT) and Evi(U). ENT retains Stream3D view selection and uses the same fusion configuration as MV-SAM3D+Stream3D. ENT yields lower geometric accuracy, while Evi(U) leads to higher perceptual errors. These results support the use of product-form evidence weighting for multi-view fusion. Third, we evaluate memory depth D. Increasing D from 1 to 3 or 5 generally improves reconstruction quality by retaining multiple high-evidence historical views. However, further increasing D to 7 does not consistently help. Overall, $D = 5$ provides a practical trade-off between reconstruction quality, appearance, and memory cost. Finally, we analyze TopK view selection. Using too few views ($K = 4$) reduces multi-view coverage and degrades performance. Once enough informative views are selected, performance becomes relatively stable: increasing K from 8 to 16 yields only moderate gains. We therefore use $K = 8$ as a practical default, balancing quality and efficiency. Additional efficiency and memory analyses are provided in Appendix B.2.

5 Conclusion

We introduced Stream3D, a training-free framework that extends frozen view-conditioned 3D generators to long monocular streams. Stream3D uses the generator's cross-attention maps to identify reliable historical conditioning views for each 3D query token. This evidence is stored in a compact Adaptive Evidential Memory and used by Evidence-Based Multi-Generation to select a bounded view bundle for each chunk. As a result, Stream3D keeps evidential memory constant with stream length while leaving the original generator unchanged. Experiments on GSO and NAVI show that Stream3D improves both appearance and geometry over single-view generators, multi-view diffusion baselines, and streaming alternatives such as KV-cache reuse and flow-based feature editing. Limitations are provided in the appendix.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, structural revision, LaTeX formatting, and critical review of methodological claims and evaluation procedures. All AI-assisted outputs were reviewed and verified by the authors, who take responsibility for the final content, including its text, claims, and artifacts.

REFERENCES

3DTopia. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023.

Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12608–12618, 2023.

Anthropic. Introducing Claude Opus 5, 2026. URL https://www.anthropic.com/news/claude-opus-5.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7996–8006, 2024.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 2023.

Carlos Campos, Richard Elvira, Juan J. Gomez Rodriguez, Jose M. M. Montiel, and Juan D. Tardos. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, et al. SAM 3: Segment anything with concepts, 2025.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22246–22256, 2023.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025a.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025b.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024a.

Ziwen Chen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Fuxin Li, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint arXiv:2410.12781*, 2024b.

Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1052–1067, 2007.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2553–2560. IEEE, 2022.

Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pp. 834–849. Springer, 2014.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018.

Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pp. 15–22. IEEE, 2014.

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *Advances in Neural Information Processing Systems*, volume 35, pp. 31841–31854, 2022.

Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and rgb-d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21584–21593, 2024.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16860–16870, 2025.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *Advances in Neural Information Processing Systems*, 36:76061–76084, 2023.

Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. Consistent4d: Consistent 360 dynamic object generation from monocular video. In *International Conference on Learning Representations*, 2024.

Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023. doi: 10.1145/3592433. Article 139.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. In *Advances in Neural Information Processing Systems*, 2024.

Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pp. 225–234. IEEE, 2007.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Escher-Net: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9503–9513, 2024.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19721–19730, 2025.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024a.

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024b.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, volume 2026, pp. 141261–141285, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023c.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16651–16662, 2025.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18039–18048, 2024.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12663–12673, 2023.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pp. 405–421, 2020.

Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pp. 2320–2327. IEEE, 2011.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9914–9925, 2024.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung W Kim, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023b.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation. In *International Conference on Machine Learning*, pp. 31915–31929, 2023.

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15598–15607, 2021.

Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024a.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. DreamGaussian: Generative Gaussian splatting for efficient 3D content creation. In *International Conference on Learning Representations*, 2024b.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22819–22829, 2023.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12619–12629, 2023a.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jing Wang, Fengzhuo Zhang, Xiaoli Li, Vincent YF Tan, Tianyu Pang, Chao Du, Aixin Sun, and Zhuoran Yang. Error analyses of auto-regressive video diffusion models: A unified framework. *arXiv preprint arXiv:2503.10704*, 2025b.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024a.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023c.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024b.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality meshes. *arXiv preprint arXiv:2404.12385*, 2024.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024a.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024b.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025b.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024a.

Yinghao Xu, Zifan Shi, Yifan Wang, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024b.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6796–6807, 2024.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Karsten Kreis, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023.

Kaichen Zhou, Jia-Wang Bian, Jian-Qing Zheng, Jiaxing Zhong, Qian Xie, Niki Trigoni, and Andrew Markham. Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes. *IEEE Robotics and Automation Letters*, 2025.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, et al. Gem-4d: Geometry-enhanced video world models for robot manipulation. *arXiv preprint arXiv:2605.22882*, 2026a.

Kaichen Zhou, Yuhan Wang, Grace Chen, Gaspard Beaudouin, Fangneng Zhan, Paul Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for vggt-4d perception. In *International Conference on Learning Representations*, volume 2026, pp. 36401–36414, 2026b.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.

A Implementation Details

Evidence Score and View Selection. Entropy computation uses 10^{-10} for numerical stability, and confidence is clipped to $[0, 1]$. Cache updates merge entries by global frame index and retain the top-D unique frames per query, breaking score ties by frame index. The float32 score and int64 frame-index arrays occupy $12QD$ bytes, corresponding to 240 KiB for $Q = 4096$ and $D = 5$.

Multi-View Fusion. The fusion normalizer applies nonnegative clipping and the configured exponent, with uniform weights when the sum is at most 10^{-10}. Optional uniform blending and minimum-weight clipping precede final normalization. Pose-specific fields retain the first view's prediction. For SAM3D, SS and SLAT each use 25 flow steps. The SS memory and fusion-weight probes use one step at attention layer 9. SLAT candidates are probed on the current SS sparse structure at step 0 of layer 6, using the same initialization across batches of at most eight views; fusion weights are recomputed at the same step and layer. Both stages use the zero-indexed image-patch key slice $[5, 1374)$. The fusion exponent is 1, uniform blending is 0, and the minimum weights are 0.001 for SS and 0.01 for SLAT, followed by renormalization. No GT geometry, depth, or poses enter generation in the main experiments; GT poses are used only for the sensitivity analysis in Sec. B.5. Controlled comparisons use matched random-seed settings and cache budgets. For GSO, all objects and methods reset the generation random seed to the same value, 42, at every chunk. Frozen probe initialization and generation randomness are controlled separately. All baselines use the same depth predictions and camera pose estimates.

Algorithm 1 Streaming inference with Adaptive Evidential Memory.

Require: frozen generator f_θ, stream chunks $\{\mathcal{C}_k\}_{k\geq 0}$, SS/SLAT attention layers $L_{\mathrm{SS}}, L_{\mathrm{SLAT}}$, query-token range $[a, b]$, memory depth D, bundle size K, score coefficient β.

Sample $z_0 \sim \mathcal{N}(0, I)$ ▷ frozen prior for evidence probing
Initialize $\mathcal{M}[q, j] \leftarrow -\infty$ and $\mathcal{F}[q, j] \leftarrow -1$ for all (q, j)
for $k = 0, 1, 2, \ldots$ **do**
  Receive current chunk $\mathcal{C}_k$
  Let $\mathcal{U}_k$ be the views in $\mathcal{C}_k$ not previously probed
  Run one batched warmup step from z_0 for $\mathcal{U}_k$, with independent per-view conditioning
  Extract $\{\mathbf{A}_v\}_{v\in\mathcal{U}_k}$ from layer L_{SS} via Eq. 1
  Compute $\mathbf{m}_v \in [0, 1]^Q$ for each $v \in \mathcal{U}_k$ via Eq. 3
  Update $(\mathcal{M}, \mathcal{F})$ by row-wise top-D merging over query tokens
  Compute frame ownership counts $\{n_f\}$ from $\mathcal{F}$ via Eq. 4
  Select the SS bundle $\mathcal{V}^\star_{k,\mathrm{SS}}$ from the top-K frames
  Reset the generation random seed to the same fixed value at every chunk and initialize z_k
  Generate the SS sparse structure from $\mathcal{V}^\star_{k,\mathrm{SS}}$ with Eq. 5
  Probe observed views on the SS structure at layer L_{SLAT} and score via Eq. 3
  Select the SLAT bundle $\mathcal{V}^\star_{k,\mathrm{SLAT}}$ by per-latent winning-view counts
  Generate $\hat{y}_k$ from $\mathcal{V}^\star_{k,\mathrm{SLAT}}$ with Eq. 5
  Output $\hat{y}_k$
end for

Evaluation. We use the same 30-object GSO30 subset as EscherNet and MV-SAM3D, with one video stream per object and 3,000 frames in total. The NAVI subset comprises 31 objects and 31 videos, with 4,274 source frames in total. The evaluator computes the squared-L2 Chamfer distance, defined as the sum of bidirectional mean squared distances. Alpha-mask IoU uses rendered foreground masks with alpha greater than zero. PSNR, SSIM, and VGG-based LPIPS use images composited onto white with these masks. Image FID uses Inception features. With the PointNet++ backend, P-FID (Nichol et al., 2022) measures the Fréchet distance between Gaussian distributions fitted to global features of aligned generated and ground-truth point clouds. Except for the checkpoint analysis in Tab. 4, reported aggregate metrics use the final chunk of each stream. Scalar metrics are averaged across objects, while FID and P-FID use the image and point-cloud features collected from these final outputs for each method. We align every method's output to its GT 3D model using global Sim(3) registration and ICP refinement before computing metrics. The resulting scores assess geometry and surface appearance up to global similarity, not absolute pose or metric scale.

A.1 Implementation of Controlled Baselines

The view-selection baselines in Tab. 2 share the same input streams, candidate pool, chunk and bundle sizes, downstream MV-SAM3D generator, and evaluation protocol. At chunk c, the candidate pool contains all frames observed through that chunk, from which each baseline selects $K = 8$ views. The fusion and sampler variants below change only their stated components.

Random. We uniformly sample eight frames without replacement from the candidate pool and apply standard MV-SAM3D fusion.

Visibility. This baseline ranks historical views by geometric coverage using DA3 predictions. We back-project each candidate's depth map into 3D with its estimated camera pose, transform the points to the current chunk's coordinate frame, and count points that fall within the current chunk's camera frustum with valid depth. The eight views with the highest visible-point counts are passed to MV-SAM3D. The score does not use generator features or cross-attention maps.

VLM. We use Claude Opus 5 (Anthropic, 2026) to score each historical view paired with the middle frame of the current chunk. The prompt is "Given the current view of the object, how informative is this historical view for reconstructing the object? Answer with a single score from 1 to 10." We rank candidates by their scores and pass the top eight views to the same MV-SAM3D fusion. The VLM is used only for view selection.

MV-SAM3D + Stream3D. This variant uses the same selected views as Stream3D but applies MV-SAM3D's standard fusion in place of evidence-weighted velocity fusion. It isolates the contribution of evidential view selection from that of evidence-weighted fusion.

Ours (Fast). This variant retains Stream3D's Adaptive Evidential Memory, view selection, and evidence-weighted fusion, while replacing the default stage-1 sampler with the official few-step distilled sampler. It measures the latency–quality trade-off of this sampler within the same framework.

B Additional Analysis

B.1 Long-Horizon Behavior

Tab. 4 tracks reconstruction quality as more frames are observed. CD, IoU, and P-FID describe geometric quality, while image FID evaluates novel-view renderings. Geometry improves mainly during chunks 0–4. From chunk 4 onward, CD and IoU remain nearly constant, while P-FID and image FID continue to decrease. This trend suggests that later observations refine geometric detail and rendered appearance after the global shape has formed.

Fig. 6 illustrates the progression for one stream. The initially incomplete shape becomes more complete by chunk 4. The outputs at chunks 8 and 12 retain its overall geometry while surface markings become more defined. Overall, the reported checkpoints show early gains in geometric quality followed by stable global structure and continued appearance refinement.

Fig. 1 complements these checkpoint analyses with four object streams extracted from real-world videos in DL3DV-10K (Ling et al., 2024). The car, statue, tiger, and cannon streams contain 344, 354, 307, and 311 input frames, respectively. We use SAM 3 (Carion et al., 2025) to segment and track each target object across the video, then apply the same preprocessing and chunk-wise generation pipeline as on GSO and NAVI. These real-capture streams include substantial object occlusion and irregular viewpoint changes, providing a challenging setting for long-range evidence accumulation. Across these streams, Stream3D uses accumulated views to recover more complete geometry and more coherent surface appearance than single-view SAM3D.

B.2 Efficiency and Memory Footprint

Stream3D adds one batched, one-step warmup call per chunk to extract independent cross-attention evidence from its newly observed views. For a 100-frame sequence with $K = 8$, generation takes

Table 4: **Long-horizon reconstruction quality on GSO.** SAM3D uses one randomly selected observed view at each horizon.

Metric	Method	Chunk index (Input frames seen)								
		0 (8)	1 (16)	2 (24)	3 (32)	4 (40)	5 (48)	6 (56)	7 (64)	12 (100)
CD-L2↓	SAM3D	0.0961	0.0928	0.0949	0.0935	0.0952	0.0920	0.0944	0.0932	0.0940
	Stream3D	0.0872	0.0766	0.0644	0.0486	0.0484	0.0482	0.0481	0.0480	0.0480
IoU↑	SAM3D	0.655	0.670	0.661	0.669	0.660	0.667	0.663	0.666	0.664
	Stream3D	0.677	0.705	0.741	0.770	0.772	0.773	0.774	0.774	0.775
P-FID↓	SAM3D	72.9	70.4	71.8	70.9	72.3	70.1	71.6	71.0	71.263
	Stream3D	63.6	52.4	48.2	42.5	42.1	41.6	41.2	41.0	40.641
Image FID↓	SAM3D	107.4	103.6	106.0	104.1	106.8	103.1	105.6	104.7	105.197
	Stream3D	95.9	88.0	87.6	88.9	83.4	77.0	73.6	71.9	66.711

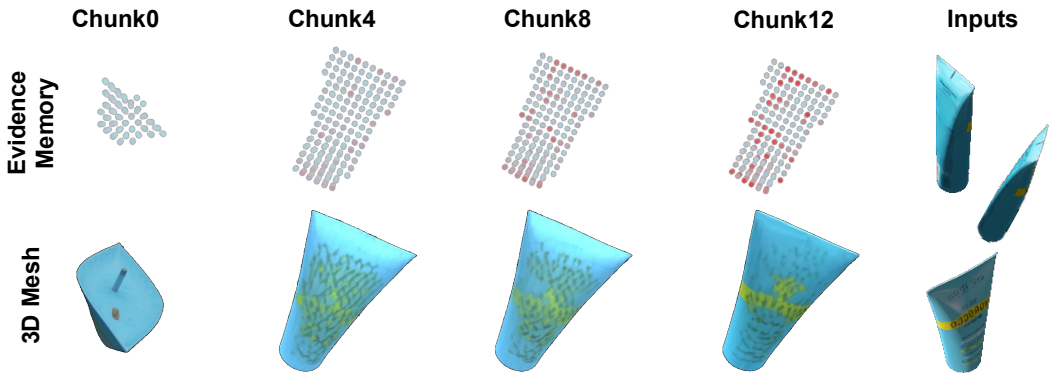


Figure 6: **Qualitative temporal consistency of Stream3D.** Top: evidential memory at selected chunks of a single stream, with blue-to-red indicating earlier-to-later observations. Bottom: the corresponding 3D reconstructions. Input views are shown on the right.

955 s for Stream3D and 945 s for MV-SAM3D, a 1.1% difference (Tab. 5). The reported total sequence times are 1062 s and 945 s, respectively, a 12.4% difference after including warmup and other measured overhead. The persistent evidential memory is lightweight. As shown in Tab. 6, it stores only the top-D evidence scores and frame indices for each query token. Its size scales as $O(QD)$ and remains constant with respect to stream length. In practice, reconstructing the selected bundle also requires the RGB frames and metadata, such as poses and depths, which are read on demand from external storage. The persistent cache therefore remains fixed in size with respect to stream length.

Tab. 7 compares generation time alone as K increases from 4 to 8. Stream3D rises from 649 to 955 s, while MV-SAM3D rises from 638 to 945 s. Stream3D-Fast uses the same memory, selection, and fusion mechanism as Stream3D, but replaces the default stage-1 sampler with the official few-step distilled sampler. It therefore represents a latency–quality trade-off within the same framework rather than a separate method.

B.3 Evidence-Score Fusion

Tab. 8 compares evidence-weighting variants for multi-view fusion. Our default product form combines attention mass and spatial concentration. ENT retains the same selected views and uses the fusion configuration of MV-SAM3D+Stream3D. Relative to ENT, the product variant improves CD from 0.053 to 0.0478 and Image FID from 77.75 to 66.71. The sum and unnormalized variants also yield lower accuracy than the product form under the tested settings.

Table 5: **Efficiency.** Runtime for a 100-frame sequence with $K = 8$ on a single H100. Generation total excludes warmup.

Model	Warmup / chunk (s)	Generation total (s)	Total / sequence (s)	Per frame (s)	Peak GPU (GB)
Stream3D	6.9	955	1062	10.6	18.4
Stream3D-Fast	**6.7**	**486**	**590**	**5.9**	**18.3**
MV-SAM3D	n/a	945	945	9.45	18.4

Table 6: **Evidence memory footprint.** Storage for float32 scores and int64 frame indices with $Q = 4096$.

D	Evidence memory	Shape
1	48 KiB	4096×2
2	96 KiB	4096×4
3	144 KiB	4096×6
4	192 KiB	4096×8
5	240 KiB	4096×10

Table 7: Generation time for a 100-frame sequence at two bundle sizes K.

Model	$K = 4$ gen total (s)	$K = 8$ gen total (s)	$K = 8$ / $K = 4$
Stream3D	649	955	1.47×
MV-SAM3D	638	945	1.48×

Table 8: Ablation on evidence-score fusion variants.

Variant	CD-L2 ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Img-FID ↓
Product (default)	**0.0478**	**0.775**	**40.64**	**16.145**	**0.866**	**0.1396**	**66.71**
Sum $\alpha = 0.25$	0.0505	0.7681	44.21	15.82	0.8603	0.1481	75.88
Sum $\alpha = 0.50$	0.0527	0.7591	44.36	15.64	0.8586	0.1500	76.61
Sum $\alpha = 0.75$	0.0545	0.7536	44.42	15.55	0.8584	0.1521	77.67
Evi(N)	0.056	0.749	44.98	15.730	0.863	0.1486	74.08
Ent	0.053	0.760	47.26	15.892	0.864	0.1449	77.75

B.4 Random-Seed Sensitivity

We repeat Stream3D(SAM3D) on GSO-30 with seeds 41, 42, and 43, keeping all other settings fixed and resetting the selected seed at every chunk. Tab. 9 reports the final-chunk results. CD ranges from 0.0478 to 0.0492, and Image FID from 65.93 to 67.84; the reported seed 42 remains close to the three-seed mean across all metrics. The observed variation is limited under these tested seeds.

Table 9: Seed sensitivity of Stream3D(SAM3D) on GSO-30. Seed 42 is the setting reported in the main paper.

Seed	CD-L2↓	IoU↑	P-FID↓	PSNR↑	SSIM↑	LPIPS↓	Img-FID↓
41	0.0492	0.7731	41.17	16.082	0.8648	0.1412	67.84
42	0.0478	0.7754	40.64	16.145	0.8660	0.1396	66.71
43	0.0481	0.7760	40.09	16.192	0.8672	0.1381	65.93
Mean	0.0484	0.7748	40.63	16.140	0.8660	0.1396	66.83
Std	0.0007	0.0015	0.54	0.055	0.0012	0.0016	0.96

B.5 Pose and Pointmap Sensitivity

Tab. 10 probes the effect of geometric conditioning on GSO-30 with Stream3D(SAM3D). Removing pose and pointmap together raises CD from 0.0478 to 0.0562, reduces IoU from 0.7754 to 0.7368, and lowers PSNR by 1.015 dB. Replacing DA3 pose with GT pose while retaining DA3 depth gives smaller changes: CD decreases to 0.0463 and PSNR rises to 16.34 dB. The GT-pose setting is used only for this sensitivity analysis. These results show a benefit from geometric conditioning and a limited further gain from GT pose under the tested protocol; the joint-removal setting does not isolate the contributions of pose and pointmap.

Table 10: Sensitivity to DA3 pointmap and pose on GSO-30.

Setting	CD-L2↓	IoU↑	P-FID↓	PSNR↑	SSIM↑	LPIPS↓	Img-FID↓
DA3 pose + DA3 depth (default)	0.0478	0.7754	40.64	16.145	0.8660	0.1396	66.71
w/o pose & pointmap	0.0562	0.7368	46.83	15.13	0.8521	0.1538	74.92
GT pose + DA3 depth	0.0463	0.7782	39.78	16.34	0.8681	0.1374	65.24

B.6 More Visulization

Fig. 7 compares ground truth, SAM3D, and Stream3D on four objects. SAM3D processes individual observations independently, while Stream3D uses informative views retained from the stream.

Fig. 8 further compares Stream3D with single-view and multi-view generation baselines. Single-view methods are constrained by partial observation and must hallucinate large unobserved regions. Multi-view diffusion baselines improve consistency by fusing multiple views, but they operate on a bounded view set and lack an explicit long-range memory mechanism. Stream3D improves over these baselines by retaining token-level evidence from the stream and selecting the most informative historical observations for generation.

Fig. 9 compares different streaming strategies. Cache- and transport-based methods reuse latent or feature states, but these states can become stale or accumulate errors over long sequences. Frame-level MV-SAM3D variants are more stable, but selecting views globally can discard observations that are important for specific spatial regions. Stream3D addresses this by maintaining token-wise evidential memory, which enables different query tokens to rely on different historical views while keeping the conditioning bundle bounded.

C Limitations

The method is built on top of the quality of the underlying pretrained generator. If the base model fails to reconstruct the object from individual views, the proposed evidential memory cannot fully recover the missing geometry or appearance.



Figure 7: **Qualitative comparison between SAM3D and Stream3D.** For each object, we show ground truth, the reconstruction from Stream3D, and the single-view SAM3D result.



Figure 8: **Qualitative comparison with single-view and multi-view 3D generation baselines.** We compare the ground truth with SAM3D, TRELLIS+M.D., TRELLIS.2, TRELLIS.2+M.D., Stream3D(SAM3D), and Stream3D(TRELLIS.2). Single-view methods recover plausible visible regions but often hallucinate incomplete or inconsistent unseen geometry. Multi-view diffusion improves view consistency with a bounded input set, but can still miss fine geometry or produce artifacts under long-stream observations. Stream3D improves both completeness and consistency by selecting informative historical views through token-level evidential memory, and the gains are visible across both SAM3D and TRELLIS.2 backbones.

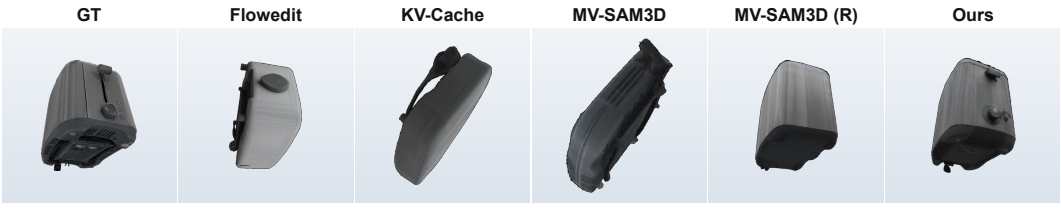


Figure 9: **Qualitative comparison with streaming baselines.** We compare the ground truth with FlowEdit, KV-Cache, MV-SAM3D, MV-SAM3D with random view selection, and Stream3D(SAM3D). FlowEdit improves short-range consistency but is limited by local latent editing and can lose long-range evidence. KV-Cache carries historical states forward but does not explicitly filter them by geometric reliability. MV-SAM3D variants select a compact set of frames at the view level, which can miss local evidence needed by specific spatial regions. In contrast, Stream3D retains token-level evidence over time and selects a bounded conditioning bundle, leading to more complete geometry and more stable appearance under long-stream generation.